Chapter 5 – Sampling from infinite width networks

Qing Yang

Deep Learning Lecture Notes Reading @ CUHKSZ

Feb 23, 2021

Outline

- Lemma 1: Maurey in Hilbert spaces
- Lemma 2: Maurey for signed measures

Sampling in Hilbert space

- First consider sampling in Hilbert spaces.
- Suppose $X = \mathbb{E}V$, where r.v. V is supported on a set S.
  $\widehat{X} = \frac{1}{k}\sum_{i=1}^{k} V_i$, where $(V_1, \cdots, V_k)$ are sampled i.i.d.
- Want to argue $\hat{X} \approx X$, i.e. try to make the Hilbert norm $\|X - \widehat{X}\|$ small.
- Maurey Lemma.

Lemma 1

(Maurey) Let $X = \mathbb{E}V$ *be given, with* V *supported on* S*, and let* $(V_1, \cdots, V_k)$ *be i.i.d draws from the same distribution. Then*

$$\mathbb{E}_{V_1,\cdots,V_k} \left\| X - \frac{1}{k}\sum_{i=1}^{k} V_i \right\|^2 \le \frac{\mathbb{E}\|V\|^2}{k} \le \frac{\sup_{U\in S}\|U\|^2}{k},$$

and moreover there exist $(U_1, \cdots, U_k)$ *in* S *so that*

$$\left\| X - \frac{1}{k}\sum_{i=1}^{k} U_i \right\|^2 \le \mathbb{E}_{V_1,\cdots,V_k} \left\| X - \frac{1}{k}\sum_{i=1}^{k} V_i \right\|^2 .$$

Proof of Lemma 1

Let $(V_1, \cdots, V_k)$ be i.i.d as stated.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{V_1,\ldots,V_k} \left\| X - \frac{1}{k}\sum_i V_i \right\|^2 = \mathop{\mathbb{E}}_{V_1,\ldots,V_k} \left\| \frac{1}{k}\sum_i (V_i - X) \right\|^2 \\
&= \mathop{\mathbb{E}}_{V_1,\ldots,V_k} \frac{1}{k^2} \left[\sum_i \|V_i - X\|^2 + \sum_{i \neq j} \langle V_i - X, V_j - X \rangle \right] \\
&= \mathop{\mathbb{E}}_{V} \frac{1}{k} \|V - X\|^2 \qquad \text{(EV=X, i. i. d)} \\
&= \mathop{\mathbb{E}}_{V} \frac{1}{k} \left(\|V\|^2 - \|X\|^2 \right) \qquad \text{(E<V, X>=||X||^2)} \\
&\leq \mathop{\mathbb{E}}_{V} \frac{1}{k} \|V\|^2 \leq \sup_{U \in S} \frac{1}{k} \|U\|^2 .
\end{aligned}
$$

And there must exist $(U_1, \cdots, U_k)$ in S so that

$$
\left\| X - \frac{1}{k}\sum_i U_i \right\|^2 \leq \mathbb{E}_{V_1,\cdots,V_k} \left\| X - \frac{1}{k}\sum_i V_i \right\|^2 .
$$

Apply Lemma 1 to infinite-width networks

Recall Definition 2.1: An infinite-width shallow network is characterized by a signed measure ν over weight vectors in $\mathbb{R}^p$:

$$x \longmapsto \int \sigma(w^T x) d\nu(w).$$

Issue:

- what is the appropriate Hilbert space?
  We will use $\langle f, g \rangle = \int f(x)g(x)dP(x)$ for some probability measure P on x, so $\|f\|^2_{L_2(P)} = \int f(x)^2 dP(x)$.
- our "distribution" on weights may not be a probability.
  E.g. consider $x \in [0, 1]$ and $\sin(2\pi x) = \int_0^1 1[x \geq b] 2\pi \cos(2\pi b) db$ (univariate approximation). There are two issues

  $$\int_0^1 |2\pi \cos(2\pi b)| db \neq 1,$$

  and $\cos(2\pi b)$ takes on negative and positive values.

Correction

Write a generalized shallow network as

$$x \longmapsto \int g(x; w) d\mu(w),$$

where μ is a nonzero signed measure over some abstract parameter space $\mathbb{R}^p$. E.g., $w = (a, b, v)$ and $g(x; w) = a\sigma(v^T x + b)$.

- Decompose $\mu = \mu_+ - \mu_-$ into nonnegative measures $\mu_\pm$ with disjoint support (Jordan decomposition).
- For nonnegative measures, define total mass $\|\mu_\pm\|_1 = \mu_\pm(\mathbb{R}^p)$, and otherwise $\|\mu\|_1 = \|\mu_+\|_1 + \|\mu_-\|_1$.
- Define $\widetilde{\mu}$ to sample $s \in \{\pm 1\}$ with probability

$$\mathbb{P}[s = +1] = \frac{\|\mu_+\|_1}{\|\mu\|_1}, \quad \mathbb{P}[s = -1] = \frac{\|\mu_-\|_1}{\|\mu\|_1},$$

  and then sample $g \sim \frac{\mu_s}{\|\mu_s\|_1} \triangleq \widetilde{\mu}_s$ and output $\widetilde{g}(x; w, s) = s\|\mu\|_1 g(x; w)$.

One property after correction

This sampling procedure has the correct mean.

$$\int g(x;w)d\mu(w) = \int g(x;w)d\mu_+(w) - \int g(x;w)d\mu_-(w)$$

$$= \quad \|\mu_+\|_1 \mathbb{E}_{\widetilde{\mu}_+} g(x;w) - \|\mu_-\|_1 \mathbb{E}_{\widetilde{\mu}_-} g(x;w) \quad \left(\frac{\mu_s}{\|\mu_s\|_1} \triangleq \widetilde{\mu}_s\right)$$

$$= \quad \|\mu\|_1 \left[\mathbb{P}_{\widetilde{\mu}}[s=+1]\mathbb{E}_{\widetilde{\mu}_+} g(x;w) - \mathbb{P}_{\widetilde{\mu}}[s=-1]\mathbb{E}_{\widetilde{\mu}_-} g(x;w)\right]$$

$$\left(\mathbb{P}\big[s=\pm 1\big] = \frac{\|\mu_\pm\|_1}{\|\mu\|_1}\right)$$

$$= \quad \mathbb{E}_{\widetilde{\mu}} \widetilde{g}(x;w,s)$$

Maurey for signed measures

Lemma 2

(Maurey for signed measures) Let μ denote a nonzero signed measure supported on $S \subseteq \mathbb{R}^p$ and write $g(x) := \int g(x; w) d\mu(x)$. Let $(\tilde{w}_1, \cdots, \tilde{w}_k)$ be i.i.d draws from the corresponding $\widetilde{\mu}$ and let P be a probability measure on x. Then

$$\mathop{\mathbb{E}}_{\tilde{w}_1,\ldots,\tilde{w}_k} \left\| g - \frac{1}{k} \sum_i \tilde{g}(\cdot; \tilde{w}_i) \right\|_{L_2(P)}^2 \leq \frac{\mathbb{E} \left\| \tilde{g}(\cdot; \tilde{w}) \right\|_{L_2(P)}^2}{k}$$
$$\leq \frac{\|\mu\|_1^2 \sup_{w \in S} \|g(\cdot; w)\|_{L_2(P)}^2}{k},$$

and moreover there exist $(w_1, \cdots, w_k)$ in S and $s \in \{\pm 1\}^k$ with

$$\left\| g - \frac{1}{k} \sum_i \tilde{g}(\cdot; w_i, s_i) \right\|_{L_2(P)}^2 \leq \mathop{\mathbb{E}}_{\tilde{w}_1,\ldots,\tilde{w}_k} \left\| g - \frac{1}{k} \sum_i \tilde{g}(\cdot; \tilde{w}_i) \right\|_{L_2(P)}^2 .$$

Proof of Lemma 2

By the property above (mean not change), we know that $g = \mathbb{E}_{\widetilde{\mu}} \widetilde{g}$, where $\widetilde{g} = s\|\mu\| g(\cdot; w)$. So by the regular Maurey (Lemma 1) applied to $\widetilde{\mu}$ and Hilbert space $L_2(P)$, i.e. $V = \widetilde{g}$ and $g = \mathbb{E} V$,

$$\left(\textit{recall } \mathbb{E}_{V_1, \cdots, V_k} \left\| X - \frac{1}{k} \sum_{i=1}^{k} V_i \right\|^2 \leq \frac{\mathbb{E}\|V\|^2}{k} \leq \frac{\sup_{U \in S} \|U\|^2}{k}, \right)$$

$$\begin{aligned} \mathop{\mathbb{E}}_{\tilde{w}_1, \ldots, \tilde{w}_k} \left\| g - \frac{1}{k} \sum_i \tilde{g}(\cdot; \tilde{w}_i) \right\|_{L_2(P)}^2 &\leq \frac{\mathbb{E} \|\tilde{g}(\cdot; \tilde{w})\|_{L_2(P)}^2}{k} \\ &\leq \frac{\sup_{s \in \{\pm 1\}} \sup_{w \in \mathcal{W}} \| \|\mu\|_1 s g(\cdot; w) \|_{L_2(P)}^2}{k}. \\ &\leq \frac{\|\mu\|_1^2 \sup_{w \in S} \|g(\cdot; w)\|_{L_2(P)}^2}{k}, \end{aligned}$$

and the existence of the fixed (w_i, s_i) is also from Maurey.

An example (various infinite-width sampling bounds)

- Suppose $x \in [0,1]$ and f is differentiable. Using our old univariate calculation,

$$f(x) - f(0) = \int_0^1 \mathbf{1}[x \geq b] f'(b) \mathrm{d}b.$$

Let μ denote $f'(b)db$, then a sample $((b_i, s_i))_{i=1}^k$ from $\widetilde{\mu}$ satisfies

$$\left\| f(\cdot) - f(0) - \frac{1}{k} \sum_i s_i 1[\cdot \geq b_i] \|\mu\|_1 \right\|_{L_2(P)}^2$$

$$\leq \quad \frac{\|\mu\|_1^2 \sup_{b \in [0,1]} \|1[\cdot \geq b_i]\|_{L_2(P)}^2}{k} = \frac{1}{k} \left(\int_0^1 |f'(b)| db \right)^2.$$

Recall $\widetilde{g} = s\|\mu\| g(\cdot; w)$ and Lemma 2:

$$\left\| g - \frac{1}{k} \sum_i \tilde{g}(\cdot; w_i, s_i) \right\|_{L_2(P)}^2 \leq \mathop{\mathbb{E}}_{\tilde{w}_1, \ldots, \tilde{w}_k} \left\| g - \frac{1}{k} \sum_i \tilde{g}(\cdot; \tilde{w}_i) \right\|_{L_2(P)}^2$$

$$\mathop{\mathbb{E}}_{\tilde{w}_1, \ldots, \tilde{w}_k} \left\| g - \frac{1}{k} \sum_i \tilde{g}(\cdot; \tilde{w}_i) \right\|_{L_2(P)}^2 \leq \frac{\mathbb{E} \|\tilde{g}(\cdot; \tilde{w})\|_{L_2(P)}^2}{k}$$

$$\leq \frac{\|\mu\|_1^2 \sup_{w \in S} \|g(\cdot; w)\|_{L_2(P)}^2}{k}$$

Example (continued)

- Now consider the Fourier representation via Barron's theorem:

$$f(x) - f(0) = -2\pi \int \int_0^{\|w\|} 1[w^T x - b \geq 0] \left[\sin(2\pi b + 2\pi\theta(w))|\hat{f}(w)|\right] dw$$

$$+2\pi \int \int_{-\|w\|}^{0} 1[-w^T x + b \geq 0] \left[\sin(2\pi b + 2\pi\theta(w))|\hat{f}(w)|\right] dw,$$

where $\hat{f}(w)$ is the Fourier transform:

$$\hat{f}(w) := \int \exp(-2\pi i w^T x) f(x) dx.$$

Then Maurey's lemma implies that there exist $((w_i, b_i, s_i))_{i=1}^k$ such that, for any probability measure P support on $\|x\| \leq 1$,

$$\left\| f(\cdot) - f(0) - \frac{1}{k} \sum_i s_i \mathbf{1}[\langle w_i, \cdot \rangle \geq b_i] \right\|_{L_2(P)}^2$$

$$\leq \frac{\|\mu\|_1^2 \sup_{w,b} \|\mathbf{1}[\langle w, \cdot \rangle \geq b]\|_{L_2(P)}^2}{k}$$

$$\leq \frac{4\|\widehat{\nabla f}(w)\|^2}{k}.$$

Here $\widehat{\nabla f}(w) = -2\pi i w \hat{f}(w)$, whereby $\|\widehat{\nabla f}(w)\| = 2\pi \|w\| \cdot |\hat{f}(w)|$.

Derive Fourier representation via Barron's theorem (Chapter 4)

Fourier inversion:

$$f(x) := \int \exp(2\pi i w^T x) \hat{f}(w) dw.$$

Write $\hat{f}(w) = |\hat{f}(w)| \exp(2\pi i \theta(w))$ with $|\theta(w)| \leq 1$. Since f is real-valued,

$$\begin{aligned}
f(x) &= \Re \int \exp(2\pi i w^\intercal x) \hat{f}(w) \mathrm{d}w \\
&= \int \Re \left(\exp(2\pi i w^\intercal x) \exp(2\pi i \theta(w)) |\hat{f}(w)| \right) \mathrm{d}w \\
&= \int \Re \left(\exp(2\pi i w^\intercal x + 2\pi i \theta(w)) \right) |\hat{f}(w)| \mathrm{d}w \\
&= \int \cos \left(2\pi w^\intercal x + 2\pi \theta(w) \right) |\hat{f}(w)| \mathrm{d}w.
\end{aligned}$$

Then

$$
\begin{aligned}
& f(x) - f(0) \\
&= \int \left[\cos(2\pi w^\top x + 2\pi\theta(w)) - \cos(2\pi w^\top 0 + 2\pi\theta(w))\right] \boxed{|\hat{f}(w)|} \mathrm{d}w
\end{aligned}
$$

Using that $\|x\| \leq 1$,

$$
\begin{aligned}
& \cos(2\pi w^\top x + 2\pi\theta(w)) - \cos(2\pi\theta(w)) \\
&= \int_0^{w^\top x} -2\pi \sin(2\pi b + 2\pi\theta(w)) \mathrm{d}b \\
&= -2\pi \int_0^{\|w\|} \mathbf{1}[w^\top x - b \geq 0] \sin(2\pi b + 2\pi\theta(w)) \mathrm{d}b. \\
&+ 2\pi \int_{-\|w\|}^{0} \mathbf{1}[-w^\top x + b \geq 0] \sin(2\pi b + 2\pi\theta(w)) \mathrm{d}b.
\end{aligned}
$$

Plugging this into the previous form,

$$f(x) - f(0) = -2\pi \int\int_0^{\|w\|} \mathbf{1}[w^\intercal x - b \geq 0] \left[\sin(2\pi b + 2\pi\theta(w))|\hat{f}(w)|\right] \mathrm{d}b\mathrm{d}w,$$
$$+ 2\pi \int\int_{-\|w\|}^{0} \mathbf{1}[-w^\intercal x + b \geq 0] \left[\sin(2\pi b + 2\pi\theta(w))|\hat{f}(w)|\right] \mathrm{d}b\mathrm{d}w,$$

and the integral can be upper bounded with

$$\begin{aligned}
&\left|2\pi \int \int_0^{\|w\|} \mathbf{1}[w^\intercal x - b \geq 0] \left[\sin(2\pi b + 2\pi\theta(w))|\hat{f}(w)|\right] \mathrm{d}b\mathrm{d}w\right| \\
&+ \left|2\pi \int\int_{-\|w\|}^{0} \mathbf{1}[-w^\intercal x + b \geq 0] \left[\sin(2\pi b + 2\pi\theta(w))|\hat{f}(w)|\right] \mathrm{d}b\mathrm{d}w\right| \\
&\leq 2\pi \int \int_{-\|w\|}^{\|w\|} |\sin(2\pi b + 2\pi\theta(w))|\,|\hat{f}(w)|\mathrm{d}b\mathrm{d}w \\
&\leq 2\pi \int 2\|w\| \cdot |\hat{f}(w)|\mathrm{d}b\mathrm{d}w \\
&= 2 \int \left\|\widehat{\nabla f}(w)\right\| \mathrm{d}w.
\end{aligned}$$

This upper bounds the mass of the weight distribution.


Thank you